Managing Unstructured Data With Structured Legacy Systems

David A. Maluf
david.a.maluf@nasa.gov

Peter B. Tran
peter.b.tran@nasa.gov
NASA Ames Research Center
Intelligent Systems Division
Mail Stop 269-4
Moffett Field, CA 94035

Abstract—In this paper we describe an approach and system for managing and joining enterprise semi-structured data in a high-throughput, nimble, and scalable systems with traditional relational database management systems (RDBMS). This paper presents the second release of NASA's NETMARK system. NETMARK is an Enterprise Information Integration (EII) framework based on a modern "schema-less" concept approach. NETMARK "schema-less" information integration reinvents the way of managing semi-structured documents within traditional RDBMS. We describe in particular detail the unique underlying data storage approach and efficient query processing mechanisms given the new proposed storage system upgrade. We present an extensive evaluation of the virtual union between NETMARK with the persistent schemas similar to commercial off-the-self products, such as Systems Applications and Products (SAP), currently utilized for NASA's Financial System, through well validated applications. At the heart of the approach is the philosophy of a well-defined and focused approach on most common data management requirements in the enterprise, and not burdening users and application developers with unnecessary complexity and formal data integration processes. This paper presents the details of achieving the integration between two incompatible systems. [1] [2]

TABLE OF CONTENTS

1. INTRODUCTION

The Problem of Heavy Middleware: The 1990's witnessed a significant amount of activities towards the development of EII (Enterprise Information Integration) technologies that were aimed at addressing the ubiquitous problem of providing data integration across multiple, distributed and possibly heterogeneous data sources in the enterprise. Software vendors that included both start-up companies, as well as larger players, such as International Business Machine (IBM) were offering software based upon a "middleware" architecture. The idea behind this particular architecture was that the middleware would provide an integrated access layer across the various data sources being integrated. This is integration as opposed to the data warehousing approach in which all data is loaded and centralized at a single place for further analysis. Towards the late 90's, eXtensible Markup Language (XML) gained prevalence simultaneously addressing the problem of syntactic heterogeneity but not *semantic* heterogeneity across different information sources.

The functional key issue with the middleware architecture approach to seamless data integration is the significant amount of effort and resources required for managing and reconciling heterogeneous data schemas. Schemas describing data in the individual information sources, as well as specifying linkages across multiple schemas, to form an integrated view of the information. The amount of time and resources required for schema management becomes a key impediment to EII technology being scalable and cost-effective for large applications. Indeed as observed in an EII technology review [1] *"A connected thread to this (key impediments for EII) is to address modeling and metadata management, which is the highest cost item in the first place"*. The original vision of intelligent information integration to nimble data management for an integrated access to information sources on-demand went awry. We trace this to some tacit, incorrect assumptions regarding how enterprise data should be managed and integrated. These assumptions, along with our alternative approach to addressing these issues are:

[1] 1-4244-1488-1/08/$25.00 ©2008 IEEE.
[2] IEEEAC paper #1360, Version 2, Updated November 08, 2007

1. Data must always be stored and managed in DBMS systems. Actually, the requirements of applications vary greatly, ranging from simple data that can be stored in simple document-oriented formats (e.g. spreadsheets and/or flat text files) to complex large data objects (e.g. binary image files) that does indeed require DBMS storage.

2. The database must always provide for and manage the structure and semantics of the data through formal schemas. Alternatively, the "database" can be nothing more than intelligent storage. Data could be stored generically and imposition of structure and semantics (schema) may be done by clients as needed.

3. Managing multiple schemas from several independent data sources and their interrelationships between them is inevitable and unavoidable; thus produces "schema-chaos". Alternatively, any imposition of schema can be done by the clients, only as when needed by applications.

2. NETMARK AS AN EXTENSIBLE DATABASE

Middleware technology should be a *cost-effective* solution and not become part of the problem as it is now. At the NASA Ames Research Center, we have designed and developed a data management and integration system called NETMARK that achieves data integration across multiple structured and unstructured data sources in a highly scalable and cost efficient manner. The querying and integration of originally unstructured data, such as various formatted reports in Microsoft Word, Adobe Portable Document Format (PDF), Excel spreadsheets, and PowerPoint presentations, is a key focus, given that the bulk of enterprise data is indeed unstructured. A new paradigm we introduce is that of *context-sensitive* querying and search. Let us illustrate this with some examples. Consider a document report in Microsoft Word format (see Figure 1). The document comprises of several text 'fragments' (e.g. sections and sub-sections, such as the section titled "Abstract", "Project Summary", "Background", etc.) A spreadsheet in Excel can also be fragmented into various rows, columns, cells, tables, and workbooks (or sets thereof). Similarly, a PowerPoint slide comprises typically of a slide title and some slide content. Each such sections or sub-sections are considered as topic headings to a particular *context.* For instance, in the Word report we have a 'Background' context, in the example PowerPoint slide we may have a 'Constellation Spirals' context as a heading, etc. The actual document content then becomes the text, graphics, or other material within the document fragment that is then referred to as *content*. For instance, the text in the 'Background' paragraph heading is the content associated with that particular context.

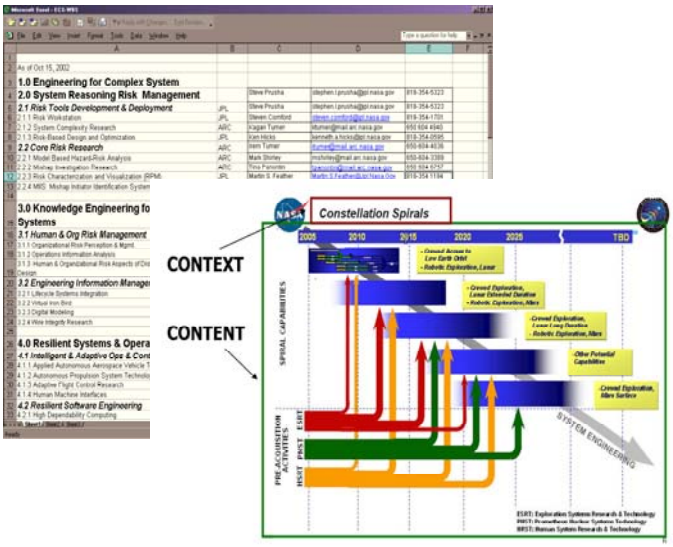

Figure 1: Fragments in unstructured data

Users pose queries in terms of context and content where they are able to search and retrieve particular textual fragments of interest. For instance, a query, such as "Context=Procurement" [3] would return all fragments from a collection of documents, where the context contains the word 'Procurement' (case insensitive). Similarly, the query "Context=Procurement & Content=Contract" will return all fragments which contains the keyword word 'contract' *within the context of* 'Procurement'.

3. A LEAN APPROACH

NETMARK supports such context and content oriented queries over a collection of unstructured data of literally any common types found in the enterprise. This has proven to be a powerful and effective paradigm for information retrieval in real-world applications. In addition to data management and integration, we have also considered other key issues in the enterprise information lifecycle. Providing data, to an integrated application, should be an easy process requiring minimal effort from the user. In fact many existing EII technologies require that any data to be integrated should be massaged, parsed, or marked to be a certain format or "wrapped" for translation. NETMARK provides a capability where data can simply be provided as-is. Providers simply drag and drop their data (e.g. a folder with several reports, spreadsheets, etc. sitting on the user desktop) into a "NETMARK-enabled" server. While maintaining the simple concept of the folder on their desktop, the NETMARK system than formats and structures it appropriately for seamless integration. At the data consumer end, we further provide capabilities for quickly composing reports and presentations over the integrated data. Finally, the NETMARK system incorporates and interoperates with open and widely used data representation and exchange standards and protocols. All data is ultimately represented

[3] We illustrate using an informal query syntax here

and stored in XML formats using open protocols such as the Web Distributed Authoring and Versioning (WebDAV)[4] are used for client-server communications. WebDAV is a W3C (World Wide Web Consortium) Internet Engineering Task Force (IETF) Request for Comments (RFC) standard that provides a set of extensions in the forms of methods, headers, and content-type ancillaries to the HTTP/1.1 protocol for resource management, namespace manipulation, and locking mechanisms for collaboratively sharing and editing documents remotely from web-enabled HTTP servers. Originally, RFC2518 was established in February 1999 and has been superseded by RFC4918 as of June 2007. The implementation however is purely relational, meeting the ultimate objective of integration with legacy RDBMS systems.

4. NETMARK

The NETMARK architecture is illustrated in Figure 2.

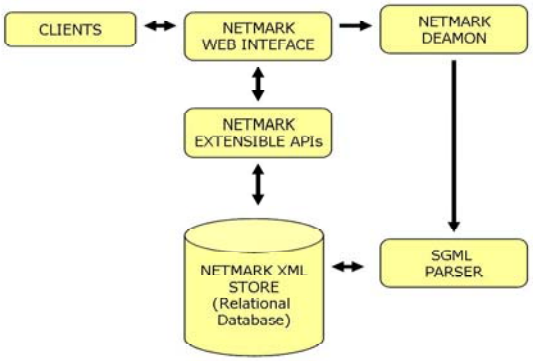

Figure 2: NETMARK Architecture

Various clients, such as data producers and providers and data consumers (or both), access NETMARK through a web interface. We will provide some illustrations of this interface on both context and content querying. The NETMARK daemon processes the incoming client requests and the Standard Generalized Markup Language (SGML)[5] parser provides functionality for loading data (e.g. documents) into NETMARK. The SGML parser hands the superset of conventional Hyper Text Markup Language (HTML) web pages and XML data formats. SGML is the precursor and superset of markup language known as International Standards Organization (ISO) 8879 standard specification on "Information Processing for Text and Office Systems" published in 1986. The NETMARK daemon is a continual process that reads in any new documents inserted into a NETMARK folder via drag and drop features and then invokes the customizable SGML parser for structuring it and loading it into the NETMARK XML data store. The data store is a relational DBMS that has been optimized to handle any data structure using a "schema-less" paradigm concept.

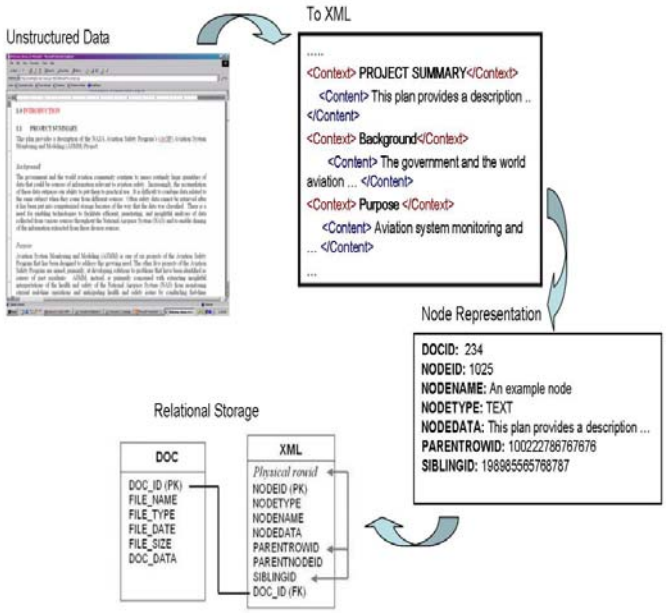

Figure 3: Data storage in NETMARK

Data storage in NETMARK, as shown in Figure 3, has two the relational tables (e.g. DOC and XML which map objects to relations. A document could then be joined by a primary (PK) and foreign key (FK) pair named "DOC_ID". The XML table is recursively nested on physical row identifiers called a "ROWID". A ROWID is a pseudo-column within the relational tables that allow the fastest single block read access to the tabular record. The DOC table contains the document metadata fields, such as file name, type, date, and size, etc.

The "schema-less" concept demonstrates the ability to articulate context independently from content reflecting on structures, such as tags or attributes of an XML entry, but in a dynamic fashion. The context and content oriented manner in which all data is modeled leads to a very efficient mechanism for storing the data coupled with efficient retrieval. The data storage "cycle" is as follows:

Unstructured data is provided to NETMARK by placing the data in a NETMARK folder. NETMARK then automatically structures the data and converts it to XML. This conversion is done based on heuristics that takes into account the document format (e.g. titles, headings etc.) to fragment the document. Each document gets marked up as context and content blocks of XML fragments. Each block is then represented as a *node* in a hierarchical document tree-like data structure. We will not go into the details of a node later; but a node is essentially the fundamental unit that captures the information in each context and content.

[4] WebDAV (RFC2518) specifications as of February 1999: http://www.webdav.org/specs/rfc2518.html
WEBDAV (RFC4918) specifications as of June 2007: http://www.ietf.org/rfc/rfc4918.txt

[5] SGML specification (ISO8879): http://www.w3.org/TR/html401/intro/sgmltut.html

The nodes are stored in a relational table called "XML". The information or metadata about the document is also stored and maintained as well in the second table called "DOC". Note that with this data representation strategy, the information in any document is ultimately stored in the same two relational tables XML and DOC. This representation is independent of any "schema" associated with the document and is thus termed to be "schema-less."

5. APPLICATION

A key feature of the NETMARK storage system is the simplicity by which applications and users can manage data and retrieve data efficiently unlike X-Path and X-Search based systems which relies on complex query languages. The challenges for new developers and non-technical users to adopt this system are not well-suited to ad-hoc queries of information in a semi-structured data store such as NETMARK. NETMARK provides the ease of use of a full-text search engine with the capabilities of a semi-structured data store for information retrieval and analysis. NETMARK has been deployed in several applications in the NASA enterprise. It serves as the integration engine for other more expansive systems for information and process management. With NETMARK we have been able to assemble new integration applications very quickly with minimal software development effort (zero in many cases) and typically requiring just about 2 man days for system setup and application assembly. One such application is the analysis of mishap reports for Aviation Safety within NASA. Such analysis reports are typically text-based reports describing the analysis of a range of accidents involving government (both NASA and other agencies) and commercial aircrafts. Using NETMARK, we are able to select particular sections and sub-sections from multiple reports and further load this information to data analysis and visualization tools seamlessly. The integrated access and analysis capabilities over integrated data have proved invaluable to Aviation Safety Analysts at NASA. Also, the assembly of the application was done with minimal development effort and time.

NETMARK success has many industry spillovers leading to commercialization and collaboration efforts. For example, the "NASA-Xerox" (NX) system is the result of a strategic collaboration between NASA and XEROX Corporation, where NETMARK has been integrated with many XEROX DocuShare[6] enterprise content management systems with capabilities for text and document management. NX offers a suite of capabilities in:

1. Content management, including capabilities for document management and collaborative sharing tool;
2. Content process management, for business process activities and management (BPM), such as action item tracking, workflow activities, and compliance.

NASA has implemented the NX technology at several NASA centers and in various NASA missions and programs, including the following:

1. The NASA International Space Station (ISS) uses NX to mine information for historical decisions and safety assurance information;
2. NASA Program Analysis and Evaluation (PA&E) adopted the commercial version of NX in 2005, which led to adoption by the NASA's Strategic Management Council (SMC);
3. Most NASA centers use the NX platform, including Ames Research Center (ARC), Langley Research Center (LaRC), Dryden Flight Research Center (DFRC), Jet Propulsion Laboratory (JPL), Johnson Space Center (JSC), and NASA Headquarters (HQ).

The ability of integrating information is well demonstrated with a complete turn-key application known as the Program Management Tool (PMT). PMT is a custom-built business intelligence solution for NASA to successfully manage large programs and projects. PMT integrates over nine distinct data sources with periodical data updates. NETMARK is the underlying data management and integration engine for PMT. PMT generates financial data and integrates with the eminent data streams from NASA's Business Warehouse (BW) SAP-based system. PMT enables program, project, and task managers to communicate successfully any critical information on the status and progress of all program levels in an efficient and update-to-date manner. PMT keeps track of program and project goals, risks, milestones and deliverables, and assists the proper allocation of financial, material, and human resources. It is well integrated with other agency-wide information systems, such as BW. PMT supports all essential program and project management activities and corresponding documents, such as the creation and monitoring of annual task plans, monthly reporting of technical, schedule, management, budget status, tracking budget phasing plans, analyzing program risks, and mitigation strategies. It will assist in reporting and evaluating project life cycle costs, accessing convenient aggregated views, and automatically creating "Earned Value Management" (EVM) assessments, Quad-Charts and other analytical reports. PMT also provides integrated access to multiple distributed resources across the NASA agency. Some the NASA agency-wide information systems PMT has interfaced with are, namely the "ERASMUS" reporting system (ERASMUS is an executive reporting system and project performance dashboard that includes performance metrics of all NASA centers, programs, projects, and safety

[6] XEROX DocuShare is a registered trademark of XEROX Corporation. http://docushare.xerox.com/
All other trademarks are of their respective companies.

and health activities), the NASA Technology Inventory Database (an inventory of technologies developed by or under development at NASA), and the Integrated Financial Management System (IFMP is an agency-wide information system supporting NASA financial management activities) thereby significantly reducing cost and time for entering the same data multiple times into different systems. This overall reduces data redundancies in multiple information systems. NETMARK has received the "*Best Practices in Storage Award*" by Computer World and Storage Networking World magazine publication.

6. CONCLUSIONS

While NETMARK exists for government and commercial use for few years, the NETMARK development team's intention to release the NETMARK system with an open source initiative for limited availability sometimes in the near future. The type of open source license for NETMARK and its capabilities, policies, and processes for receiving the software has yet to be determined; but NASA Ames Research Center has a history of releasing software into the open source community and we expect that the NETMARK system will be released in this manner also. Interested individuals and/or groups may contact Dr. David A. Maluf at David.A.Maluf@nasa.gov for additional information. Also, additional information about the NASA's Program Management Tool (PMT), including system overview, demonstrations, and documentations, can be obtained at the following website: http://pmt.arc.nasa.gov

REFERENCES

[1] A. Halevy, N. Ashish, D. Bitton, M. Carey, D. Draper, J. Pollock, A. Rosenthal and V. Sikka, "Enterprise information integration: Successes, challenges and controversies ," in *ACM SIGMOD*, 2005

[2] D. Maluf, D. Bell, N. Ashish, C. Knight and P. Tran, "Semi-structured data management in the enterprise: A nimble, high-throughput, and scalable approach," in *International Database Engineering and Applications Symposium (IDEAS)*, 2005

[3] Maluf, A. David, Bell, David, Ashish, Naveen, "Lean Middleware, '' Proceedings of the 2005 ACM SIGMOD international conference on Management of data. 2005, pp. 788–791, ISBN:1-59593-060-4, 2005.

[4] Maluf, A. David, Bell, David, "NASA Program Management Tool," *Project Management Challenge, University of Maryland*, 2005.

[5] Maluf, A. David, Bell, David, "Towards G2G: Systems of Technology Database Systems, " *IEEE Aerospace Conference*, Montana, 2005.

[6] Maluf, David A., Gawdiak, Yuri, Bell, David, "On Space Exploration and Human Error: A Paper on Reliability and Safety," *Proceedings of the Thirty-Eighth Annual Hawaii International Conference on System Sciences*, 2005.

[7] Maluf, David A., Tran, Peter B., "NETMARK: A Schemal-Less Extension for Relational Databases for Managing Semi-Structured Data Dynamically," *International Symposium on Methodologies for Intelligent Systems, Lecture Notes in Computer Science, Springer Verlag*, 2003.

[8] Maluf, David A., Tran, Peter B, "NETMARK: Adding Hierarchical Object to Relational Databases with Schema-less Extensions ", *Intelligent Systems Design and Applications (ISDA)*, Tulsa, Oklahoma, Conference Proceedings, 2003.

[9] Maluf, A. David, Bell, G. David, Knight, Chris, Tran, Peter, La, Tracy, Lin, Jenessa, McDermott, Bill, Pell, Barney,"*NASA-XDB-IPG: Extensible Database - Information Grid" Global Grid Forum 8, 2003.*

BIOGRAPHIES

David A. Maluf *received his Ph.D. from McGill University in 1995 and his postdoctoral from Stanford University. He has been involved in Intelligent Information Integration and databases since. David was also Director of Software Development at Incyte. Before NASA, David founded and operated Science Gate as CTO. The company was successfully acquired. At NASA, David was the Project Manager for Knowledge Engineering under the Engineering for Complex Systems program. David was the CIO for the program. In conjunction with the FAA, David has been leading, from its inception, the development and operation of large government information grid projects, connecting US government centers nation wide. David is the inventor on many NASA patents, including Netmark tool suites, which were commercialized leading to products such as NX and PMT. David is the recipient of many NASA Awards: Best Technology Commercialization, Turning Goals into Reality, and Space Act Awards.*

Peter B. Tran *is currently a Senior IT Software Architect for NASA Ames Research Center working on data integration and information management projects for the NASA's Constellation Program. Previously, Peter worked as a software consultant, architect, technical lead, and software engineer at several technology companies, including QSS Group, Inc., BEA Systems, XUMA, Computer Sciences Corporation, and Recom Technologies. Peter has a degree in electrical engineering and computer sciences from the University of California at Davis, and has taken graduate-level coursework at Stanford University majoring in Computer Science.*